Scientific Research Publishing

Journal of Software Engineering and Applications, 2019, 12, 35-50
http://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Prediction of Criminal Suspects Based on Association Rules and Tag Clustering

Bo Cheng[1], **Weihong Li**[1*], **Haoxin Tong**[2]

[1]School of Geography, South China Normal University, Guangzhou, China
[2]R & D Department, Aerospace Finest (Guangdong) Information Technology Co. Ltd., Guangzhou, China
Email: 15625110283@163.com, *hongweili9981@163.com

How to cite this paper: Cheng, B., Li, W.H. and Tong, H.X. (2019) Prediction of Criminal Suspects Based on Association Rules and Tag Clustering. *Journal of Software Engineering and Applications*, **12**, 35-50.
https://doi.org/10.4236/jsea.2019.123003

Received: February 26, 2019
Accepted: March 24, 2019
Published: March 27, 2019

Abstract

To date, not many studies have been conducted on criminal prediction. In this study, the criminal data related to city S is divided into a training data set and a validation data set at a 1:1 ratio in light of the personal tag data and the travel and accommodation data of criminals and ordinary people in city S. Firstly, the FP-growth algorithm is adopted to calculate association rules between the criminals and the ordinary people in their travel and hotel accommodation data, in order to discover criminal suspects based on association rules. Secondly, the DBSCAN algorithm is employed for clustering of the tag data of the criminals and the ordinary people, followed by similarity calculation, in order to discover criminal suspects based on tag clustering. Lastly, intersection operation is performed on the above two sets of criminal suspects, and the resulting intersection is verified against the criminal validation set for elimination of criminals who appear in the intersection so as to obtain final criminal suspects. Results show that a set of 648 criminal suspects is retrieved based on the association rules calculated by the FP-growth algorithm, while a set of 973 criminal suspects is retrieved based on DBSCAN clustering and cosine similarity of the personal tags; the number of criminal suspects is narrowed down to 567 after the intersection operation of the two sets, and 419 of the 567 criminal suspects are further verified to be criminals using the validation set, thereby leaving the other 148 to be the final criminal suspects and giving a prediction accuracy of 73.9%. The data mining method of criminal suspects based on association rules and tag clustering in this study has been successfully applied to the police system of city S, and the experiment proves the effectiveness of this method in detecting criminal suspects.

Keywords

FP-Growth, Association Rule, DBSCAN, Tag Clustering, Criminal Suspects

1. Introduction

Nowadays, crime situation is becoming increasingly serious across the globe with more crime types and a higher number of criminals, posing a threat to human lives and property as well as social stability. Public security authorities are increasingly tasked with maintaining public security and fighting crimes with an ever-growing requirement on law enforcement. Given the constantly generated crime data, it is necessary for data analysts to reveal hidden patterns in the data, analyze implicit relationships between the data, predict occurrence of crimes and discover potential criminals, so as to improve the efficiency of law enforcement efficiency of public security authorities and prevent occurrence of crimes.

Association rule mining (ARM) [1] is a research focus in the field of data mining. As one of the classic algorithms for data mining, ARM has been widely used in crime research. It is possible to extract relevant criminal evidence from association rules of a large number of data items, and further explore the patterns, trends and links between different crimes, so as to provide support for the police in case investigation and crime prevention. ARM performs well in exploring the causes of crime, identifying the main crime suspects, and having a deeper insight into crime series. A variety of ARM algorithms have been playing an important role in crime analysis and crime prediction. In the international research community of association-rule-based crime mining, Ng *et al.* [2] introduced temporal association rules and proposed an incremental algorithm to solve the problem of how to process time series whose association rules contain time expressions, and employed the new algorithm to discover crime patterns in Hong Kong. Buczak *et al.* [3] explored applying fuzzy association rule mining to recognition of community crime patterns, and such application promoted the efficiency of local law enforcement. Tan *et al.* [4] were the first to analyze the role of the FP-growth algorithm in computer crime forensics, specifying the limitation that the FP-growth algorithm had when used to discover the latest crimes and serious crimes, and making some improvements to the FP-growth algorithm and its test. Joshi *et al.* [5] proposed the FP-Tree similarity algorithm and found it more effective than the Apriori algorithm. Usha *et al.* [6] [7] tested the Apriori algorithm and other algorithms such as the Fp-growth algorithm in real and synthetic crime data sets, and found that each algorithm had its own advantages. Shekhar *et al.* [8] explored spatial frequent pattern mining (SFPM) based on criminal pattern analysis (CPA) and validated this mining method with a spatial crime dataset. Isafiade *et al.* [9] revisited frequent pattern growth models for crime pattern mining, proposing a revised frequent pattern growth (RFPG) model, and also proposed a descriptive statistical approach based on a quartile (floor-ceil) function, which was used to identify recurring trends in criminal activity. Based on geographic and demographic factors, Asmai *et al.* [10] used ARM to generate a crime mapping model for crime analysis, employing the model to examine the occurrence of crime at a specific location, and demonstrating that the model could be used to analyze future crime locations with relatively high crime potential and improve crime prevention implementation. Extensive research has been conducted in China on ARM-based crime mining. Based on fuzzy set and rough set theory, Chen and Yu [11] [12] employed ARM to quantitatively analyze a criminal dataset, make deductions, and extract rules, providing theoretical guidance for crime prevention. ARM has gained widespread attention and application in the fields of criminal portraits and criminal forensic analysis [13]. Moreover, ARM has been widely used in crime research such as crime investigation [14], criminal suspect analysis [15], criminal behavior analysis [16], and reoffending [17] [18]. In view of the temporal and spatial characteristics of crime data, many studies have proposed a number of improved ARM algorithms such as spatio-temporal association rules [19] [20], cluster association rules [21], and data cube-based association rules [22], as well as other improved algorithms such as incremental association rules [23].

In summary, association rules have been extensively applied to crime mining in relevant studies, the effectiveness of ARM in different fields of crime mining has been investigated in depth, and a large number of improved ARM algorithms have been proposed. These studies are mainly focused on using crime data for association rules mining, but in practice more frequently encountered is business data, which is not necessarily related to crime but is easier to obtain, thereby making it crucial to extract crime information from a large volume of ordinary business data. Moreover, existing studies are largely focused on crime case analysis and crime pattern mining instead of the mining and prediction of potential criminals, and most of the studies are focused on macro-level factors that affect the occurrence of crimes while not considering micro-level characteristics of criminals, while the micro-level characteristics are the internal factors that determine whether an individual would commit a crime. This study combines ARM algorithm and clustering algorithm to deal with crime and related data. This method is different from previous crime mining methods. It can directly find criminal suspects instead of criminal hotspots or others, and it makes full use of ordinary business data, which is easier to access and handle. In this study, ARM is performed on city S-related travel data and accommodation data of criminals and ordinary people, and meanwhile, tag clustering is performed on criminals and ordinary people, in order to discover as criminal suspects, the ordinary people who not only frequently travel with criminals but also have highly similar tags to them. Given that criminal acts are often carried out in the form of gangs, the people who are discerned to travel with the criminals and have a high similarity to the criminals in tags can be considered as potential key individuals. Monitoring criminal suspects can greatly reduce the incidence of infraction and crimes and improve public safety. In Section 2, we briefly describe the data used in the study and the preprocessing of the data. In Section 3, we provide the details of our business process modeling methodology. In Section 4, we describe the research results and analysis in detail, and test and compare the algorithm. In Section 5, we summarize the conclusions of this work.

2. Research Data

City S-related travel and accommodation data of criminals and ordinary people in 2016 as well as their personal tag data at that time are collected. The travel and accommodation data consist of shuttle ticketing data and hotel accommodation data (Table 1). A total of 600,000 and 2 million personal attribute data are collected on criminals and ordinary people, respectively, with the data consisting of personal ID numbers and relevant personal tags (Table 2). The criminal data is divided into a training data set and a validation data set at a 1:1 ratio. The training data set is used for criminal suspect-related computation, while the validation data set is used for verification of the method effectiveness by checking whether actual criminals are among the criminal suspects.

The personal attributes in Table 2, except for ID number, are treated as personnel tags, and the meanings of some personal tags are shown in Table 3.

Clustering analysis and similarity calculation are performed on vector inputs. Tag vectorization is conducted according to the publicly accessible corpus of Chinese word vectors developed by Beijing Normal University and Renmin University of China—a corpus providing pre-trained 300-dimensional (300 d) character vectors. Each tag vector is calculated according to the following formula:

$$\mathrm{V} = \frac{1}{m}\sum_{i=1}^{m} v_i$$

where V is a calculated tag vector, with m representing the number of Chinese characters in the tag and $v_i = \left(v_i^1, v_i^2, \cdots, v_i^{300}\right)$ a character vector of the tag's i-th character. Each individual has 12 tags, and therefore the tag matrix of each individual is 12 × 300 in size.

3. Research Methods

This paper first uses the FP-growth algorithm to mine the ordinary people who frequently travel with the criminals as criminal suspects based on travel data and hotel accommodation data; Then use the DBSCAN algorithm to cluster and analyze the label data of criminals and ordinary people, and obtain several tag clusters. Carry out the similarity calculation of the criminals and ordinary personnel in each cluster, and find some ordinary persons with the highest similarity with the criminals as criminal suspects; Finally, the criminal suspects based on the association rules and the criminal suspects based on the label clustering are interdigitated to obtain the final criminal suspects, and the criminals test data is used to check whether calculated criminal suspects obtain actual criminals.

Table 1. Travel accommodation data.

Data sheet	Attribute field
Shuttle	Route code, departure time, starting station, arrival station, passenger ID number
Hotel	Check-in ID number, hotel code, check-in time, hotel name

Table 2. Personal tags and meaning of the tag values.

Tag	Value	Meaning
ID number	6cbe2819c3xxxxxxxx	
Gender	M	Male
	F	Female
Age	Minor	<18 years old
	Youth	18 - 40 years old
	Middle aged	41 - 60 years old
	Elderly	>61 years old
Marital status	Married	first marriage, remarriage, and remarriage
	Unmarried	unmarried, divorced, widowed
Employment status	Employed	Currently employed
	Unemployed	Currently unemployed
Income	Low	Below the minimum wage
	Middle	Between the minimum wage and the per capita wage
	High	Higher than the per capita wage
Educational level	Primary school	Primary school education
	Junior high school	Junior high school education
	Senior high school	Senior high school education
	University	University education
Single-parent family	Yes	Raised in a single-parent family
	No	Raised in a non-single-parent family
Offspring	Yes	Having offspring
	No	Having no offspring
House property	Tenant	Having no house property
	Property owner	Having house property
Foster family	Yes	Fostered
	No	Non-fostered
Household registration	Urban	Urban household registration
	Rural	Rural household registration
Migrant	Yes	Migrant
	No	Non-migrant

The research method flow is shown in Figure 1.

3.1. Association Rule Mining Based on FP-Growth

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other

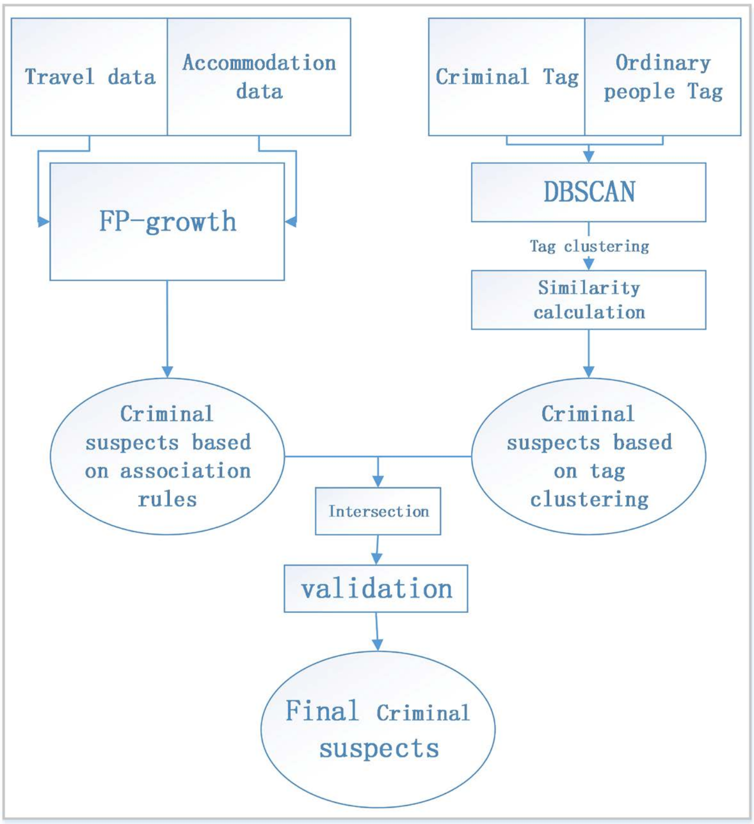

Figure 1. Research method flow chart.

forms of data repositories. Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following two parameters, the important relationships are observed:

1) Support: Indication of how frequently the itemset appears in the database. It is defined as the fraction of records that contain $X \cup Y$ to the total number of records in the database. Suppose, the support of an item is 0.1%, it means only 0.1% of the transactions contain that item.

$$\text{Support}(XY) = \text{Support count of }(XY)/\text{Total number of transaction in D}$$

2) Confidence: Fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X.

It's is a measure of strength of the association rules. Suppose, the confidence of the association rule $X \Rightarrow Y$ is 80%, it means that 80% of the transactions that contain X also contain Y together.

$$\text{Confidence}(X \mid Y) = \text{Support}(XY)/\text{Support}(X)$$

The FP-Growth Algorithm, proposed by Han in [24], is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree). In his study, Han proved that his method outperforms other popular methods for mining frequent patterns, e.g. the Apriori Algorithm [25] and the TreeProjection [26]. In some later works [27] [28] it was proved that FP-Growth has better performance than other methods, including Eclat [29] and Relim [30]. The popularity and efficiency of FP-Growth Algorithm contributes with many studies that propose variations to improve his performance.

The FP-Growth Algorithm is an alternative way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset association information.

In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity.

Based on travel data and hotel accommodation data, this study compares each car or daily hotel accommodation to a shopping basket, where the item is the ID of the person and the item set is all the people in the car or hotel. With the minimum support and confidence, the FP-growth algorithm is used to calculate the association rules between criminals and ordinary people. That is to mine ordinary people who frequently travel or stay with criminals as criminal suspects.

3.2. DBSCAN Tag Clustering and Cosine Similarity

As one of the most cited of the density-based clustering algorithms, DBSCAN is likely the best known density-based clustering algorithm in the scientific community today. The central idea behind DBSCAN and its extensions and revisions is the notion that points are assigned to the same cluster if they are density-reachable from each other. To understand this concept, we will go through the most important definitions used in DBSCAN and related algorithms. The definitions and the presented pseudo code follows the original by Ester *et al.*, but are adapted to provide a more consistent presentation with the other algorithms discussed in the paper. Clustering starts with a dataset D containing a set of points $p \in D$. Density-based algorithms need to obtain a density estimate over the data space. DBSCAN estimates the density around a point using the concept of ϵ-neighborhood.

Definition 1. ϵ **-Neighborhood.** The ϵ-neighborhood, $N(p)$, of a data point p is the set of points within a specified radius around p.

$$N\epsilon(p) = \{q \mid d(p,q) < \epsilon\}$$

where d is some distance measure and $\epsilon \in R_+$. Note that the point p is always in its own ϵ-neighborhood, *i.e.*, $p \in N\epsilon(p)$ always holds.Following this definition, the size of the neighborhood $|N\epsilon(p)|$ can be seen as a simple unnormalized kernel density estimate around p using a uniform kernel and a bandwidth of ϵ. DBSCAN uses $N\epsilon(p)$ and a threshold called minPts to detect dense regions and to classify the points in a data set into core, border, or noise points.

Definition 2. Point classes. A point $p \in D$ is classified as

- a core point if $N\epsilon(p)$ has high density, *i.e.*, $|N\epsilon(p)| \geq \text{minPts}$ where $\text{minPts} \in Z_+$ is a user-specified density threshold,
- a border point if p is not a core point, but it is in the neighborhood of a core point $q \in D$, *i.e.*, $p \in N\epsilon(q)$, or
- a noise point, otherwise.

Definition 3. Directly density-reachable. A point $q \in D$ is directly density-reachable from a point $p \in D$ with respect to and minPts if, and only if,

1) $|N\epsilon(p)| \geq \text{minPts}$, and

2) $q \in N\epsilon(p)$.

That is, p is a core point and q is in its ϵ-neighborhood.

Definition 4. Density-reachable. A point p is density-reachable from q if there exists in D an ordered sequence of points $(p_1, p_2, \cdots, p_n)$ with $q = p_1$ and $p = p_n$ such that p_{i+1} directly density-reachable from p_i $\forall i \in \{1, 2, \cdots, n-1\}$.

Definition 5. Density-connected. A point $p \in D$ is density-connected to a point $q \in D$ if there is a point $o \in D$ such that both p and q are density-reachable from o.

Definition 6. Cluster. A cluster C is a non-empty subset of D satisfying the following conditions:

1) Maximality: If $p \in C$ and q is density-reachable from p, then $q \in C$; and

2) Connectivity: $\forall p, q \in C$, p is density-connected to q.

The DBSCAN algorithm identifies all such clusters by finding all core points and expanding each to all density-reachable points. The algorithm begins with an arbitrary point p and retrieves its ϵ-neighborhood. If it is a core point then it will start a new cluster that is expanded by assigning all points in its neighborhood to the cluster. If an additional core point is found in the neighborhood, then the search is expanded to include also all points in its neighborhood. If no more core points are found in the expanded neighborhood, then the cluster is complete and the remaining points are searched to see if another core point can be found to start a new cluster. After processing all points, points which were not assigned to a cluster are considered noise.

This study uses cosine similarity to calculate the distance. Refer to the existing study [31] to set ε to 0.6 and MinPts to 255. The DBSCAN algorithm clusters the vectorized criminals and ordinary people's tag matrix, and calculates the similarity and sorting of the criminals and criminal suspects for each cluster. Then, ordinary people with similarities greater than 0.8 are selected from various clusters as criminal suspects. The cosine similarity calculation formula is as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

We use the machine learning algorithm library Spark Mllib in the distributed computing framework Spark to perform association rule calculation and cluster analysis. A total of five physical nodes are deployed for fast and efficient computing. Compared to stand-alone computing, Spark distributed computing time efficiency is increased by 300%; compared to Hadoop MapReduce distributed clusters, time efficiency is increased by 150%.

4. Results and Analysis

4.1. Analysis of Association Rule Results

Association rule computation is based on a given minimum support (min_sup) and a given minimum confidence (min_conf). A change of min_sup and min_conf will lead to a different result. The settings of min_sup and min_conf are allowed to vary according to actual needs. In this study, their settings are as follows:

$$\text{min\_sup}\left(\text{XY}\right) = 20/\text{Total number of transactions in D}$$

$$\text{min\_ conf}\left(\text{X} \mid \text{Y}\right) = 0.8$$

1) Travel association rules

For ARM of travel data, each route (including route code, departure time, start station, arrival station) is treated as a transaction ID, and all passengers on the route are treated as transaction items. Computation is performed on shuttle-related data according to a given min_sup and min_conf, and the results are expressed in the form of "route passenger X => passenger Y", which represents that when the requirements of minimum support and minimum confidence are met, passengers X and Y are considered to have taken the same route within a certain time window. However, it is unclear whether the above rules contain criminals, and therefore filtering operation is performed on all association rules to find the association rules that contain criminals. Based on each association rule that contains criminals, such as "X => Y" wherein X is assumed to be a criminal, it is possible to find all the routes that both X and Y have taken together. The ordinary individual Y who often takes the same routes as the criminal X can be considered as a criminal suspect.

The calculation results, as listed in Table 3, shows that there is a total of 433 association rules, that is, 433 criminal suspects are found. The five fields in Table 3 present the ID number of the criminal, the ID number of the criminal suspect, the confidence, the same routes that both the criminal and the criminal suspect have taken, as well as the number of the same routes taken by both the criminal and the criminal suspect—which is used as a surrogate for the support. The second rule in the table, as an example, states that the criminal with ID number 6eb9b7199bxxxxxxxx and the suspect with ID number 4ab8b81987xxxxxxxx have traveled 78 times with each other in the same routes, and the confidence of 0.9 indicates that 90% of all the routes that are taken by the criminal also involve the criminal suspect; in addition, the route information indicates that, for example,

Table 3. Travel association rules.

Criminal	Criminal suspect	Confidence	Route information	Number of the same routes
6eb9b8197cxxxxxxxxx	6eb9b8a9f8 xxxxxxxx	1.0	{652901xxxxxxxx, 2016-3-16 14:50:00, aks station, awt station}…{…}	45
6eb9bg199bxxxxxxxx	4ab8b81987 xxxxxxxx	0.9	{654025xxxxxxxx, 2016-4-26 8:45:00, dsz station, xy station}…{…}	78
6deabf198dxxxxxxxx	6deab8199c xxxxxxxx	0.8	{654221xxxxxxxx, 2016-423 10:30:00,em station, tgg station}…{…}	34
65b8b9a98gxxxxxxxx	6eb8b9a990xxxxxxxx	0.9	{654025xxxxxxxx, 2016-9-18 16:20:00, aks station, em station}…{…}	29
……	……	……	……	……
650a0ba97fxxxxxxxx	6bb8b7a992xxxxxxxx	0.8	{652829xxxxxxxx, 2016-7-26 18:10:00,xy station, zq station}…{…}	40

Note: A total of 433 association rules.

there is one route named 653124xxxxxxxx, which started at 14:50:00 on March 16, 2016 from the aks station to the awt station. As shown in the above table, different association rules have different levels of support and confidence. For example, the support and confidence in the first association rule are 45 and 1.0, respectively, indicating that the criminal 6eb9b8197cxxxxxxxxx and the criminal suspect 6eb9b8a9f8xxxxxxxx have traveled together with each other for all of the 45 routes. The second association rule shows that nearly 70 of the 78 routes that the criminal 6eb9bg199bxxxxxxxx has taken involve the criminal suspect 4ab8b81987xxxxxxxx. It is impossible yet to know which of the criminal suspects 6eb9b8a9f8xxxxxxxx and 4ab8b81987xxxxxxxx has a higher likelihood of being an actual criminal. In the third association rule, the number of the same routes and the confidence are relatively smaller compared to the first two rules, indicative of a smaller likelihood of the criminal suspect 6deab8199c xxxxxxxx being an actual criminal.

2) Hotel accommodation association rules

For each hotel, all persons who check in for accommodation on a given day are treated as one transaction. Computation is performed on the accommodation data according to given min_sup and min_conf, and the results are expressed in the form of "hotel code person X => person Y", indicating that passengers X and Y are considered to have stayed at this hotel given the fulfillment of the min_sup and min_conf requirements. By processing the hotel accommodation association rules in a similar way to the travel association rules, it is possible to find the criminal suspect Y that often stays in the same hotels as the criminal X.

Table 4 shows the analysis results of hotel accommodation association rule. There is a total of 323 association rules, that is, 323 criminal suspects are found. The information about criminals and criminal suspects staying in the same hotels is revealed in the analysis results, with the time denoting the hotel check-in

Table 4. Hotel accommodation association rules.

Criminal	Criminal suspect	Confidence	Hotel information	Times of staying in the same hotels
6edab6197fxxxxxxxx	4aaabc1980xxxxxxxx	0.9	{2016-05-01 jy inn}, {2016-05-03 xh hotel}…{…}	48
5acd29199bxxxxxxxx	1a0aa1198exxxxxxxx	0.8	{2016-05-15 fk hotel}, {2016-06-15 sd hotel}…{…}	23
6abe23198axxxxxxxx	6ab5bc1970xxxxxxxx	1	{2016-06-17 yh inn}, {2016-06-29 sy hotel}…{…}	35
654abc199dxxxxxxxx	3a0aa0198exxxxxxxx	0.8	{2016-06-28 dt inn}, {2016-07-24 xf hotel}…{…}	47
……	……	……	……	……
65280ac966xxxxxxxx	5ab5cda979xxxxxxxx	1	{2016-10-01 xh inn}, {2016-11-03 cf hotel}…{…}	24

Note: A total of 323 association rules.

date; in addition, the times of staying in the same hotels refers to the times of the criminal and the criminal suspect staying in the same hotels during the studied time window. There is a difference in support and confidence between different accommodation association rules, as is the case with the travel association rules, which indicates that different ordinary people have different likelihoods of being a criminal suspect. For example, the probability of an ordinary individual being a criminal suspect is higher in the first association rule than in the second association rule, as the former rule has higher support and confidence than the latter; the third association rule and the last association rule in Table 4 have the same confidence, but the former has a higher support, so the ordinary people in the third association rule are more likely to be a criminal suspect than in the last association rule.

Duplicates may exist between criminal suspects discovered by shuttle ticketing data and by hotel accommodation data—in other words, criminal suspects based on the association rules of shuttle ticketing may also appear in the association rules of hotel accommodation. Therefore, it is necessary to combine the two sets of results to eliminate the repetitive criminal suspects, and by dosing so a total of 648 criminal suspects has been finally obtained.

4.2. Result Analysis of DBSCAN Clustering

The results of DBSCAN clustering are shown in Table 5. Calculation gives 67 clusters, each containing a different number of people in the range of 2000 - 30,000. Each cluster contains both criminals and ordinary people, with the ratio of criminals to ordinary people varying from cluster to cluster.

Based on the tag clusters of criminals and ordinary people, the cosine similarity between ordinary people and criminals in each cluster is calculated, and ordinary people with cosine similarity greater than 0.8 are escalated as criminal suspects. The calculation results, as listed in Table 6, indicates that a total of 973 criminal suspects are obtained, with each cluster having a different number of

Table 5. DBSCAN clustering results.

Cluster code	Criminals and ordinary people	Number of the people
Cluster_1	6edab6197fxxxxxxxx,4aaabc1980 xxxxxxxx,6eb9b8a9f8 xxxxxxxx,……	2039
Cluster_2	5acd29199bxxxxxxxx,1a0aa1198e xxxxxxxx,4ab8b81987 xxxxxxxx,……	2345
Cluster_3	65b8bca9960xxxxxxxx,65cab11989xxxxxxxx,6e900aa983xxxxxxxx,……	2376
……	……	…
Cluster_66	65cab61984 xxxxxxxx,6eb8bca994 xxxxxxxx,6528011996 xxxxxxxx,……	2439
Cluster_67	6fb9b8a97cxxxxxxxx,65bcbd197dxxxxxxxx,1a0ab1198fxxxxxxxx,……	2657

Table 6. Calculation results of cosine similarity.

Cluster code	Criminal suspects	Number
Cluster_1	4b0bb21980xxxxxxxx,411abca980 xxxxxxxx,6fb9b8a968 xxxxxxxx,……	23
Cluster_2	bc08b7a969xxxxxxxx,ba11aa985 xxxxxxxx,4ab8b8a987 xxxxxxxx,……	27
Cluster_3	65b8bca9960xxxxxxxx,65cab11989xxxxxxxx,41bcb1a976 xxxxxxxx,……	15
……	……	…
Cluster_66	51aab9a987 xxxxxxxx,65bcb4a974xxxxxxxx,110aaa198fxxxxxxxx,……	17
Cluster_67	65bcbd197dxxxxxxxx,1a0ab1198fxxxxxxxx,51b9b719fexxxxxxxx,……	13

Note: A total of 973 criminal suspects.

criminal suspects—which is attributed to that each cluster is different than another in terms of both the total number of people and the number of criminals, so a cluster that is higher in both of the two numbers is also likely to have a higher number of criminal suspects identified by similarity calculation. In addition, DBSCAN clustering may result in a different compactness in a different cluster, and as a result the number of predicted criminal suspects also varies.

Moreover, the distribution of the 973 criminal suspects by their personal tags as shown in Figure 2, which shows that males are nearly twice as many as females, that is, the probability of a male individual being a criminal suspect is much higher than that of a female individual. In the distribution by age, nearly 90% of the criminal suspects are young and middle-aged, while the minors have a negligible proportion. The distribution by marital status indicates that unmarried criminal suspects are twice as many as married criminal suspects. Similarly, when it come to the distribution by employment situation, there is a huge difference between the number of employed and unemployed, with more than 80% of the criminal suspects being unemployed. In terms of household income, the low-income and middle-income people account for nearly 90% of the potential key individuals, and the number of people in each of these two income groups is nearly twice as great as that in the high-income group. Similar to income-based distribution, the education level-based distribution shows that criminal suspects with primary and junior high school education account for 80% of the potential key individuals. In addition, of the 973 criminal suspects, 65% are raised in single-parent families, 70% are raised in foster families, and more than 70% have no

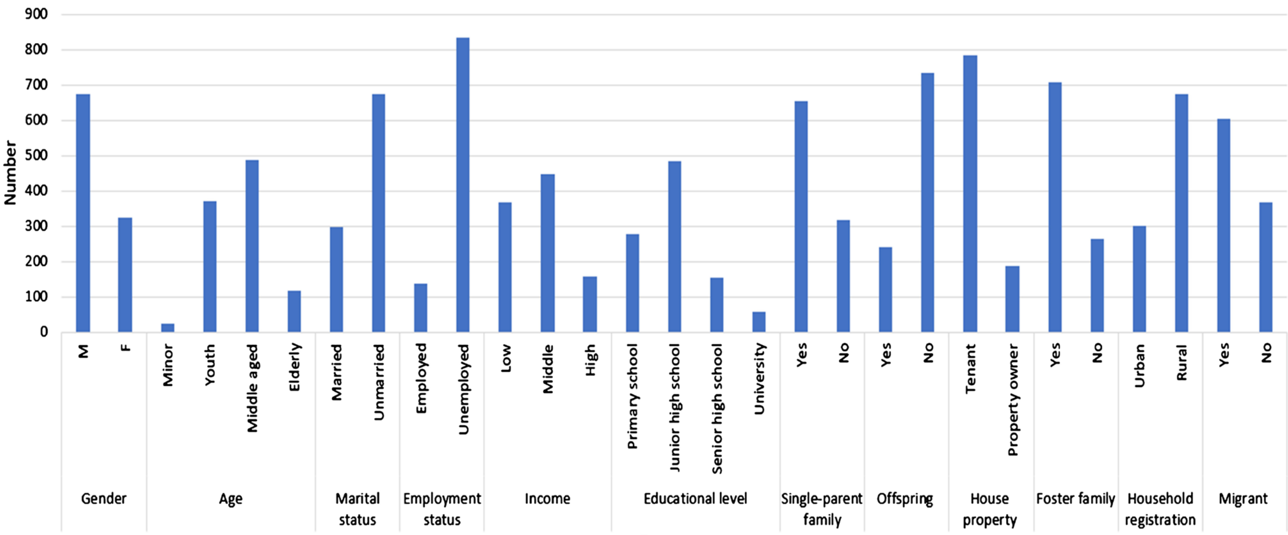

Figure 2. The distribution of potential key individuals by their personal tags.

offspring. With respect to the ownership of house property, nearly 80% of the criminal suspects do not have their own house property. The distribution by household registration and migrant status indicates that 67% of the potential key individuals have rural household registration and 60% are migrants. In summary, criminal suspects mainly have the tags of being a male, being middle-aged, being unmarried, being unemployed, having low and middle income, having received elementary school and junior high school education, being raised in a single-parent family, having no offspring, being a tenant, being raised in a foster family, having rural household registration, and being a migrant. These tags are in line with the actual situation of criminals. For example, in the comparison of a male individual versus a female individual, an employed individual versus an unemployed individual, and a low-income individual versus a high-income individual, the former is more prone than the latter to commit crimes. In summary, criminal suspects discovered by means of clustering analysis and tag similarity are in line with the tag-based distribution.

4.3. Result Validation

After ARM and DBSCAN clustering, the two sets of criminal suspects found by the two methods are subject to intersection operation. In this process, the number of criminal suspects appearing in both of the two sets is calculated to be 567, and the 567 criminal suspects are verified by the validation set of criminals, that is, these suspects are traversed in order to find the individuals appearing in the validation data—who are actual crimes, totaling 419. Therefore, with the elimination of the actual criminals, the number of criminal suspects is finally determined to be 148. We can say that the accuracy of the algorithm is 73.9%, but the actual accuracy needs to be tested in actual law enforcement activities. We can only think that the validation results here indicate the effectiveness of the above method for finding criminal suspects. Moreover, most studies predict the number of crime hotspots and crime cases, but do not predict criminal suspects. Therefore, it is difficult to compare with other crime prediction methods.

5. Conclusions

In this study, the FP-growth algorithm is used to perform ARM on travel data and hotel accommodation data, and the DBSCAN algorithm is used to achieve tag clustering of criminals and ordinary people, and finally the above results are verified. The results show that:

1) The FP-growth association rule mining algorithm shows that different association rules have different support and confidence, that is, different ordinary people have different possibilities of being criminal suspects.

2) By using the FP-growth association rule algorithm, 648 criminal suspects are found, while 973 are found through DBSCAN clustering of personnel tags; the number of criminal suspects in the intersection of the above two sets of criminal suspects is 567, and when verified against the validation data, the 567 criminal suspects are found to contain 419 actual criminals, thereby leaving 148 as the final criminal suspects with a prediction accuracy of 73.9%.

3) Criminal suspects mainly have the tags of being male, being middle-aged, being unmarried, being unemployed, having low and middle income, having received elementary school and junior high school education, being raised in a single-parent family, having no offspring, being a tenant, being raised in a foster family, having rural household registration, and being a migrant, and such tag-based distribution agrees with the situations of actual key individuals to some extent.

4) The validation results show that the method is effective in discovering criminal suspects, and it has great generalizability; that is, it can be used for data mining of train passenger information, passenger exit/entry records, Internet-café user information, as well as other travel spending information.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Agrawal, R., Imieliński, T. and Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD Record*, **22**, 207-216. https://doi.org/10.1145/170036.170072

[2] Ng, V., Chan, S., Lau, D. and Ying, C.M. (2007) Incremental Mining for Temporal Association Rules for Crime Pattern Discoveries. *Proceedings of the* 8*th Australasian Database Conference*, **63**, 123-132.

[3] Buczak, A.L. and Gifford, C.M. (2010) Fuzzy Association Rule Mining for Community Crime Pattern Discovery. *ACM SIGKDD Workshop on Intelligence and Security Informatics*, Washington DC, 25-28 July 2010, Article No. 2.

[4] Tan, Y., Qi, Z. and Wang, J. (2012) Applications of Association Rules in Computer

Crime Forensics. *Applied Mechanics & Materials*, **157-158**, 1281-1286. https://doi.org/10.4028/www.scientific.net/AMM.157-158.1281

[5] Joshi, A. and Suresh, M.B. (2014) Compact Structure of Felonious Crime Sets Using FP-Tree Comparable Algorithms Analysis. *International Journal of Computer Science and Information Technologies*, **5**, 2694-2699.

[6] Ramesh Kumar, K. and Usha, D. (2013) Frequent Pattern Mining Algorithm for Crime Dataset: An Analysis. *International Journal of Engineering Sciences & Research Technology*, **2**, 3379-3384.

[7] Usha, D. and Ramesh Kumar, K. (2014) A Complete Survey on Application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining. *International Journal of Advances in Computer Science and Technology*, **3**, 264-275.

[8] Shekhar, S., Mohan, P., Oliver, D. and Zhou, X. (2012) Crime Pattern Analysis: A Spatial Frequent Pattern Mining Approach. No. TR-12-015, Minnesota University, Minneapolis.

[9] Isafiade, O., Bagula, A. and Berman, S. (2015) A Revised Frequent Pattern Model for Crime Situation Recognition Based on Floor-Ceil Quartile Function. *Procedia Computer Science*, **55**, 251-260. https://doi.org/10.1016/j.procs.2015.07.042

[10] Asmai, S.A., Roslin, N.I.A., Abdullah, R.W., *et al.* (2014) Predictive Crime Mapping Model Using Association Rule Mining for Crime Analysis. *Science International*, **26**, 1703-1706.

[11] Chen, Z.S., Zuo, L.M. and Xia, P.P. (2008) Analysis of Extraction and Association of Characteristics of ICO Economic Crime Based on Apriori Algorithm. *Journal of Jiangxi Police Institute*, 18-22.

[12] Wang, L. and Sun, B. (2018) Analysis of Public Security Case Base Based on Data Mining. *Gansu Science and Technology*, 18-22.

[13] Wei, S.J., Zhang, J.W. and Geng, R.N. (2006) Based on the Crime Profiling Research of Computer Forensics Analysis Method. *Microcomputer Information*, **22**, 237-239.

[14] Sheng, W. (2012) Principle of Investigation Information Association Rules. *Science & Technology Information*, 300-300.

[15] Yu, N. (2015) Research on the Discovery Method of Suspicion Degree Relationship Based on the Association Rule Algorithm. Dalian Polytechnic University, Dalian.

[16] Xu, W. and Zhang, J. (2016) Research on the Association Rule Applied in the Criminality.

[17] Tang, Y.P. (2016) Recidivism Association Rule Mining Based on Apriori Algorithm. *Command Information System and Technology*, 91-95.

[18] Feng, Z.H. and Feng, Q.J. (2017) Analysis on the Features of Recidivism Based on Association Rules. *Journal of Zhejiang Sci-Tech University* (*Social Sciences Edition*), **38**, 57-60.

[19] Yan, M.Q., Guo, Z.Y. and Ren, Z.H. (2017) Spatio-Temporal Analysis of Bus Pickpocketing Using Association Rules Based on Clustering. *Journal of East China Normal University* (*Natural Science*), No. 3, 145-152.

[20] Sun, Y.H., Wang, W.J. and Chi, X.T. (2016) Correlation Mining and Prediction of Social Security Events Based on Multi-Dimensional Time Series Model. *Journal of Tianjin University* (*Social Sciences*), **18**, 97-102.

[21] Wang, H., Zheng, T. and Zhang, J.L. (2010) Cluster-Based Association Rule Algorithm in Criminal Application in Criminal Behavior Analysis. *Journal of Chinese People's Public Security University* (*Science and Technology*), **16**, 64-67.

[22] Wang, H.B., Zhang, Y.T. and Wu, S. (2016) Application of Association Rules with

Multiple Minimum Supports Based on the Data Cube in Crime Analysis. *Journal of Geomatics Science and Technology*, **33**, 405-409.

[23] Du, W. and Zhou, X.X. (2011) Research on Application of Incremental Association Rule Mining Algorithm in Criminal Behavior. *Journal of Chinese People's Public Security University* (*Science and Technology*), **17**, 56-58.

[24] Han, J., Pei, H. and Yin, Y. (2000) Mining Frequent Patterns without Candidate Generation. *Proceedings of the* 2000 *ACM SIGMOD International Conference on Management of Data*, Dallas, TX, 15-18 May 2000, 1-12. https://doi.org/10.1145/342009.335372

[25] Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules. *Proceedings of the* 20*th International Conference on Very Large Data Bases*, Santiago, Chile, 12-15 September 1994, 487-499.

[26] Agarwal, R., Aggarwal, C. and Prasad, V.V.V. (2001) A Tree Projection Algorithm for Generation of Frequent Item Sets. *Journal of Parallel and Distributed Computing*, **61**, 350-371. https://doi.org/10.1006/jpdc.2000.1693

[27] Kumar, B.S. and Rukmani, K.V. (2010) Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms. *International Journal of Advanced Networking and Applications*, **1**, 400-404.

[28] Bonchi, F. and Goethals, B. (2004) FP-Bonsai: The Art of Growing and Pruning Small FP-Trees. *Proceedings of the* 8*th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Sydney, 155-160. https://doi.org/10.1007/978-3-540-24775-3_19

[29] Zaki, M., Parthasarathy, S., Ogihara, M. and Li, W. (1997) New Algorithms for Fast Discovery of Association Rules. *Proceedings of the* 3r*d International Conference on Knowledge Discovery and Data Mining* (*KDD*'97), Newport Beach, CA, 14-17 August 1997, 283-296.

[30] Borgelt, C. (2005) Keeping Things Simple: Finding Frequent Item Sets by Recursive Elimination. *Proceedings of the* 1*st International Workshop on Open Source Data Mining*: *Frequent Pattern Mining Implementations*, Chicago, IL, 21 August 2005, 66-70. https://doi.org/10.1145/1133905.1133914

[31] Tran, T.N., Drab, K. and Daszykowski, M. (2013) Revised DBSCAN Algorithm to Cluster Data with Dense Adjacent Clusters. *Chemometrics and Intelligent Laboratory Systems*, **120**, 92-96. https://doi.org/10.1016/j.chemolab.2012.11.006